Why minimax is not that pessimistic

A. Fraysse

L2S, SUPELEC, CNRS, University Paris-Sud, 3 rue Joliot-Curie, 91190 Gif-Sur-Yvette, FRANCE. This work was performed when the author was at LTCI, Telecom ParisTech. Email: fraysse@lss.supelec.fr.

Keywords and phrases: minimax theory, maxiset theory, Besov spaces, prevalence.

1. Introduction

Since its introduction in the seventeen's, nonparametric estimation has taken a large place in the work of mathematical or signal processing communities. Often a signal has too many components, or stands in a special space, and classical estimation studies cannot be carry out. Therefore new estimation procedures, based on approximation of functions have been introduced. But which kind of estimator is the most appropriate in these cases?

This question raised a lot of definitions and discussions in the statistical community. How can two estimators be compared when they point out infinite dimensional objects and what kind of optimal behaviour can be expected. One of the most common way to test the performance of a procedure is to compare its convergence rate with an optimal one given by minimax theory. Nonetheless, this technique comes from a particular definition which can be subject to controversy. The main drawback is the pessimist point of view of this theory, which looks for the worst rate of estimation obtained in a given space. Indeed, in the minimax theory we are looking to the estimation procedure which yields the minimum of a maximum risk, in a sense to be defined, over a function space. But the worst case could be a misleading one and a method can be rejected although it is a good one for a lot of functions. The purpose of this paper is to introduce a new test of the risk, obtained thanks to genericity results. Thanks to this new kind of test we show that in fact minimax risk corresponds to a generic one.

Let us first introduce what is meant by almost every function. In a finite dimensional space, we say that a property holds almost everywhere if the

set of points where it is not true is of vanishing Lebesgue measure. The Lebesgue measure has here a preponderant role, as it is the only σ-finite and translation invariant measure. Unfortunately, no measure share those properties in infinite dimensional Banach spaces. A way to recover a natural "almost every" notion in infinite vector spaces is thus defined as follows by J. Christensen in 1972 see [2, 4, 12].

Definition 1 *Let V be a complete metric vector space. A Borel set $A \subset V$ is Haar-null (or shy) if there exists a compactly supported probability measure μ such that*

$$\forall x \in V, \quad \mu(x + A) = 0. \tag{1}$$

If this property holds, the measure μ is said to be transverse to A.

A subset of V is called Haar-null if it is contained in a Haar-null Borel set. The complement of a Haar-null set is called a prevalent set.

As it can be seen in the definition of prevalence, the main issue in proofs is to construct transverse measures to a Borel Haar-null set. We remind here two classical ways to construct such a measure.

Remark 1

1. *A finite dimensional subspace of V, P, is called a probe for a prevalent set $T \subset V$ if the Lebesgue measure on P is transverse to the complement of T.*
   This measure is not a compactly supported probability measure. However one immediately checks that this notion can be defined the same way but stated with the Lebesgue measure defined on the unit ball of P. Note that in this case, the support of the measure is included in the unit ball of a finite dimensional subspace. The compactness assumption is therefore fulfilled.

2. *If V is a function space, a probability measure on V can be defined by a random process X_t whose sample paths are almost surely in V. The condition $\mu(f + A) = 0$ means that the event $X_t - f \in A$ has probability zero. Therefore, a way to check that a property $\mathcal{P}$ holds only on a Haar-null set is to exhibit a random process X_t whose sample paths are in V and is such that*

$$\forall f \in V, \text{ a.s.} \quad X_t + f \quad \text{does not satisfy } \mathcal{P}.$$

The fact that a set is Haar-null is independent of the chosen transverse measure, as soon as the translation invariance condition is satisfied. However

this property cannot provide the exact characterization of null sets.

The following results enumerate important properties of prevalence and show that these notions supply a natural generalization of "zero measure" and "almost every" in finite-dimensional spaces, see [2, 4, 12].

Proposition 1
- *If S is Haar-null, then $\forall x \in V$, $x + S$ is Haar-null.*
- *If* $\dim(V) < \infty$, *S is Haar-null if and only if* $\mathrm{meas}(S) = 0$ *(where* meas *denotes the Lebesgue measure).*
- *Prevalent sets are dense.*
- *The intersection of a countable collection of prevalent sets is prevalent.*
- *If* $\dim(V) = \infty$, *compact subsets of V are Haar-null.*

As we can see from the properties of prevalent sets, this theory provides a natural generalization of the finite dimensional notion of almost every. Since its definition, it has been mainly used in the context of differential geometry [12] and regularity type properties [11]. A classical example is given in [11], where it is proved that the set of nowhere differentiable functions is prevalent in the space of continuous functions.Surprisingly even in the finite dimensional case, this approach has no longer been implemented. The only actual result involving genericity results in statistics is due to Doob, see [25] in the context of Bayesian estimation in parametric statistics.

Using this theory, a natural way to exhibit a test of performance for an estimating procedure is to look at the risk reached on almost every function of a function space, in the sense of prevalence.

As the minimax theory has been widely studied, a large class of results exist in different function spaces and with different risk functions. Historically, the first one is the result of Pinsker [21] which shows that suitable linear estimators reach the optimal L^2 risk rate on L^2 Sobolev classes. If the risk function is given by an L^p norm, [13, 3] show that, under certain conditions, kernel estimators are optimal in the sense of minimax theory in the same function spaces. More recent results, such as those of [20], stated that linear estimators cannot reach the optimal bound in nonlinear regression, as soon as we take the L^p risk and Sobolev classes.

In this paper we focus on Besov spaces and take the general L^p norm as a risk function. The interest of studying Besov spaces is motivated by its

practical use in approximation theory and its theoretical simplicity in terms of wavelet expansions. Furthermore, in the theoretical point of view, they also generalize some classical function space, such as Hölder and L^2 Sobolev spaces.

In those Besov spaces we study performances in terms of generic approximation of two classical estimation procedures in both white noise model and density estimation problem. With these two techniques, we will see that minimax and generic results coincide.

2. Models and estimation procedures

In the following, we consider two classical estimation problems. The first one is given by the Gaussian white noise model. Following the definition of [13], we suppose that we observe Y_t such that

$$dY_t = f(t)dt + \frac{1}{\sqrt{n}}dW_t, \quad t \in (0,1)^d, \tag{2}$$

where dW_t stands for the d-dimensional Wiener measure, n is known and f is the unknown function to be estimated.

The second theoretical framework treated in this paper is the problem of density estimation. Assume that we have access to a sequence $X_1, \ldots, X_n$ of independent and identically distributed random variables of unknown density f on $\mathbb{R}$. The problem here is to estimate f thanks to the observed sequence.

The estimation procedures that we deal with are defined thanks to a base decomposition of the function to be estimated. To define them, we first introduce the wavelet bases. In our framework, those bases allow both to define function spaces and estimation procedures. It provide thus a key tool to introduce our results. The wavelet transform is a powerful approximation tool widely used in statistics and signal processing, thanks to its properties of localization in time and frequency domains. Indeed, this property allows to reconstruct a signal with few coefficients. Its use in statistical communities and the development of wavelet based estimators are thus natural, as introduced in [18].

To define wavelets, we refer to [6] where it is proved that for r large enough there exists 2^d-1 functions $\psi^{(i)}$ with compact support and which are r regular. Furthermore each $\psi^{(i)}$ has r vanishing moments and the set of functions $\{\psi^{(i)}_{j,k}(x) = 2^{dj/2}\psi^{(i)}(2^j x - k), \quad j \in \mathbb{Z},\ k \in \mathbb{Z}^d, i \in \{1, ..., 2^d - 1\}\}$ forms an orthonormal basis of $L^2(\mathbb{R}^d)$. It is also noticed in [19] that wavelets provide unconditional bases of $L^p(\mathbb{R}^d)$ as far as $1 < p < \infty$. Taking periodized wavelets allow to restraint our studies to $[0,1]^d$.

Thus any function $f \in L^p$ can be written as

$$f(x) = \sum_{i,j,k} c^{(i)}_{j,k}\psi^{(i)}_{j,k}(x)$$

where

$$c^{(i)}_{j,k} = 2^{jd/2}\int f(x)\psi^{(i)}(2^j x - k)dx.$$

In the following we stand in isotropic cases. Thus the direction of the wavelets is not involved and for the sake of simplicity we omit the directional index i.

As the collection of $\{2^{dj/2}\psi(2^j x - k), \quad j \in \mathbb{N},\ k \in \{0, \dots, 2^j - 1\}^d\}$ form an orthonormal basis of $L^2([0,1]^d)$, observing the whole trajectory of Y_t in (2) is equivalent to treat the following problem, in which is observed $(y_{j,k})_{j\in\mathbb{N},k\in\{0,\dots,2^j-1\}^d} \in \ell^2(\mathbb{N}^{d+1})$ such that $\forall j, k$,

$$y_{j,k} = \theta_{j,k} + \frac{1}{\sqrt{n}}v_{j,k}, \tag{3}$$

where $y_{j,k} = \int \psi_{j,k}dY(x)$, v_i are i.i.d. Gaussian random variables and $(\theta_{j,k})_{j,k}$ is the sequence to be estimated.

In terms of density estimation, one can also notice that the density function to be estimated f can be represented in terms of wavelet decomposition $f = \sum \beta_{j,k}\psi_{j,k}$. In this case, the purpose is to find a sequence $(\hat{\beta}_{j,k})_{j,k}$ approximating $(\beta_{j,k})_{j,k}$.

Furthermore wavelets are useful as they provide a simple characterization of Besov spaces.

Homogeneous Besov spaces are characterized, for $p, q > 0$ and $s \in \mathbb{R}$, by

$$f \in B_p^{s,q}([0,1]^d) \iff \exists C > 0 \ \ \|f\|_{B_p^{s,q}} := \sum_{j \geq 0} \left(\sum_{k \in \{0,\dots,2^j-1\}^d} |c_{j,k}|^p 2^{(sp-d+\frac{pd}{2})j} \right)^{q/p} \leq C. \tag{4}$$

This characterization is independent from the chosen wavelet as soon as ψ has r vanishing moments, with $r \geq s$.

We also denote by $B_{p,c}^{s,q}(\mathbb{R}^d)$, the closed ball in $B_p^{s,q}(\mathbb{R}^d)$ of radius $c > 0$.

In our framework, the minimax paradigm induces that one supposes a function f belongs to $B_p^{s,q}(\mathbb{R}^d)$. Then one defines a risk or loss function thanks to a pseudo-distance on $B_p^{s,q}(\mathbb{R}^d)$, denoted $R(.,.)$. Given a radius $c > 0$ and an estimator $\widehat{f_n}$ of f which is a measurable function of the observations. In this case, the maximal risk of $\hat{f}_n$ on $B_{p,c}^{s,q}(\mathbb{R}^d)$ is defined by:

$$R^n(\hat{f}_n) = \sup_{f \in B_{p,c}^{s,q}(\mathbb{R}^d)} \mathbb{E}(R(\hat{f}_n, f)). \tag{5}$$

If $\mathcal{T}_n$ denotes the set of all measurable estimation procedures defined thanks to a given model the minimax risk on $B_{p,c}^{s,q}(\mathbb{R}^d)$ is then given by :

$$R^n(B_{p,c}^{s,q}(\mathbb{R}^d)) = \inf_{\hat{f}_n \in \mathcal{T}_n} \sup_{f \in \Theta_C} \mathbb{E}(R(\hat{f}_n, f)).$$

This minimax risk gives an optimal bound over the function class $B_{p,c}^{s,q}(\mathbb{R}^d)$. It is thus natural for estimation procedures to attempt to reach this risk, at least asymptotically when n tends to infinity.

In terms of wavelets approximation, or in any base, the most natural and classical way to define estimators is given by linear estimation.

Definition 2 *Suppose that we stand in the model* (3). *Linear estimators* $\hat{f}_n^L$ *are constructed by*

$$\hat{f}_n^L(x) = \sum_{j \geq 0} \sum_{k \in \{0,\dots,2^j-1\}^d} \hat{\theta}_{j,k}^{(n)} \psi_{j,k}(x), \tag{6}$$

where

$$\hat{\theta}_{j,k}^{(n)} = \lambda_{j,k}^{(n)} y_{j,k}.$$

Parameters $(\lambda_{j,k}^{(n)})_{j,k}$ *can be seen as smoothing weights lying in* $[0,1]$. *Those weights can be of different natures. Classical ones are:*

- *Projection weights:* $\lambda_{j,k}^{(n)} = \mathbb{1}_{\{j<T_n\}}$.
- *Pinsker weights:* $\lambda_{j,k}^{(n)} = (1 - (\frac{j}{T_n})^{\alpha})_+$,

where $(T_n)_n$ *is an increasing sequence depending on* n.

Definition 3 *Suppose that we stand in the model of density estimation. In this case, a linear estimator of the density* f *is constructed by taking*

$$\forall j \geq 0 \ \ \forall k \in \{0, \dots, 2^j - 1\}^d \quad \hat{\beta}_{j,k} = \frac{1}{n} \sum_{i=1}^{n} \psi_{j,k}(X_i). \tag{7}$$

And

$$\hat{f}_n^L(x) = \sum_{j \geq 0} \sum_{k \in \{0,\dots,2^j-1\}^d} \lambda_{j,k}^{(n)} \hat{\beta}_{j,k} \psi_{j,k}(x).$$

Where $(\lambda_{j,k}^{(n)})_{j,k}$ also are smoothing weights in $[0, 1]$.

The localization property of wavelet expansions is such that a given signal may have a sparse representation in those bases. Thus a natural estimation procedure in the white noise model, defined in [7] and ever since widely used in the signal community is to take away small wavelet coefficients. This is the principle of wavelet thresholding.

Definition 4 *Suppose that we stand in the case of white noise model* (3). *The wavelet thresholding procedure is then defined by*

$$\hat{f}_n^T(x) = \sum_{j=0}^{j(n)} \sum_{k \in \{0,\dots,2^j-1\}^d} \beta_{j,k}^T \psi_{j,k}(x). \tag{8}$$

Here the weights are given by:

$$\beta_{j,k}^T = y_{j,k} \mathbb{1}_{\{|y_{j,k}| \geq \kappa t_n\}}, \tag{9}$$

in the case of hard thresholding, or

$$\beta_{j,k}^T = \operatorname{sign}(y_{j,k})(|y_{j,k}| - \kappa t_n)_+, \tag{10}$$

for the soft thresholding. Furthermore,

$$t_n = \sqrt{\frac{\log n}{n}},$$

stands for the universal threshold and $j(n)$ is such that

$$2^{-j(n)} \leq \frac{\log n}{n} < 2^{-j(n)+1},$$

κ being a constant large enough.

Once again, in the model of density estimation wavelet thresholding is obtained thanks to a slight modification of the previous definition.

Definition 5 *Suppose that we stand in the problem of density estimation, and let $\hat{\beta}_{j,k}$ be the coefficients defined in (7). Thus the density estimator by wavelet thresholding is given by*

$$\hat{f}_n^T(x) = \sum_{j=0}^{j(n)} \sum_k \hat{\beta}_{j,k} \mathbb{1}_{\{|\hat{\beta}_{j,k}| > \kappa t_n\}} \psi_{j,k}(x).$$

Where

$$t_n = \sqrt{\frac{\log n}{n}},$$

is the universal threshold, κ is a constant large enough and $j(n)$ is such that

$$2^{-j(n)} \leq \frac{\log n}{n} < 2^{-j(n)+1}.$$

3. Statement of the main result

Let us recall minimax results in Besov spaces. Taking the L^p norm, where $1 \leq p < \infty$, as loss function, we know from [9], for our two estimation problems, that the minimax lower bound in closed balls in Besov spaces in given by the following proposition.

Proposition 2 *Let $1 \leq r \leq \infty$, $1 \leq p < \infty$ and $s > \frac{d}{r}$. Then, there exists $C > 0$ such that*

$$R^n(B_{r,c}^{s,\infty}) = \inf_{T_n} \sup_{f \in B_{r,c}^{s,\infty}} \mathbb{E}\|T_n - f\|_{L^p}^p \geq C r_n(s,r,p)$$

where

$$r_n(s,r,p) = \begin{cases} n^{-\frac{ps}{2s+d}} & \text{if } r > \frac{dp}{2s+d}, \\ \left(\frac{n}{\log n}\right)^{-\frac{p(s-\frac{d}{r}+\frac{d}{p})}{2(s-\frac{d}{r})+d}} & \text{else.} \end{cases}$$

Let us now check what is known concerning estimation procedures that we deal with. Although it is proved in [9] that thresholding procedures reach asymptotically the optimal rate up to a logarithmic correction, it is not always the case for linear procedures. As it can be seen in [8], with L^2 risk, linear estimators do not attain the minimax rate when studied functions have a sparse representation in a given base. This result is generalized by the following proposition from [8] which gives the optimal rate that can be reached in this case.

Proposition 3 *Let $1 \leq r \leq \infty$, $1 \leq p < \infty$ and $s > \frac{d}{r}$. There exist $C > 0$ such that*

$$R^n_{lin}(B^{s,\infty}_{r,c}) = \inf_{\tilde{T}_n \, linear} \sup_{f \in B^{s,\infty}_{r,c}} \mathbb{E}\|\tilde{T}_n - f\|^p_{L^p} \geq C \tilde{r}_n(s,r,p)$$

where

$$\tilde{r}_n(s,r,p) = \begin{cases} n^{-\frac{ps}{2s+d}} & \text{if } r > p \\ \left(\frac{n}{\log n}\right)^{-\frac{ps'}{2s'+d}} & \text{else,} \end{cases}$$

and $s' = s - \frac{d}{r} + \frac{d}{p}$.

We see in the following theorem that Proposition 3 remains true if we replace the risk maximum by the risk reached on almost every function. We also prove that in the same context Proposition 2 is satisfied by thresholding algorithms up to a logarithmic term. We say in the following that $a_n \approx b_n$ if $\frac{\log a_n}{\log b_n} \to 1$.

Theorem 1 *Let $1 \leq r \leq \infty$, $1 \leq p < \infty$ and $s > \frac{d}{r}$. Then, in the context of (2) or for the problem of density estimation:*

- *For almost every function f in $B^{s,\infty}_r([0,1]^d)$,*

$$\inf_{\widehat{f}^L_n \, linear} \mathbb{E}\|\hat{f}^L_n - f\|^p_{L^p} \approx n^{-\alpha p}, \tag{11}$$

  where

$$\alpha = \begin{cases} \frac{s}{2s+d} & \text{if } r \geq p, \\ \frac{s - \frac{d}{r} + \frac{d}{p}}{2(s - \frac{d}{r} + \frac{d}{p}) + d} & \text{else.} \end{cases} \tag{12}$$

- *For almost every function in $B^{s,\infty}_r([0,1]^d)$, and for thresholding estimator $\hat{f}^T_n$*

$$\mathbb{E}\|\hat{f}^T_n - f\|^p_{L^p} \approx \left(\frac{n}{\log n}\right)^{-\alpha p} \tag{13}$$

where

$$\alpha = \begin{cases} \frac{s}{2s+d} & \text{if } r > \frac{pd}{2s+d} \\ \frac{(s-\frac{d}{r}+\frac{d}{p})}{2(s-\frac{d}{r})+d} & \text{else.} \end{cases} \tag{14}$$

As mentionned earlier, this generic result, such as Doob's theorem does not provide the exact behaviour of a given function. However, it introduces a new vision of a generic behavior in Besov spaces and of the minimax theory as in these particular cases, generic and minimax results coincide.

As we will see in the following section, the proofs of these results are quite simple. They are mainly based on the maxisets theory. Once known the maxiset associated to an estimation procedure, one study the genericity of such a set in the involved function space. The advantage is that our theorem can be easily extended to another kind of estimation procedures, thanks for instance to the results of [1, 23] or to other spaces, such as Sobolev spaces.

4. Proof of Theorem 1

For the sake of completeness, let us recall some basic facts upon the maxisets theory.

4.1. Maxiset theory

The maxiset theory introduced recently in [5, 16, 17] is an alternative way to compare different estimation procedures. In our case, it provides a crucial key to prove Theorem 1. The main idea is to look for the maximal space on which an estimator will reach a given rate instead of searching an optimal rate for a given space.

Definition 6 *Let ρ be a risk function and $(v_n)_{n\in\mathbb{N}}$ a sequence such that $v_n \to 0$. For $\hat{f}_n$ an estimator, the maximal space associated to ρ, v_n and a constant T is given by*

$$MS(\hat{f}_n, \rho, v_n, T) = \left\{ f;\ \sup_n v_n^{-1} \mathbb{E}(\rho(\hat{f}_n, f)) < T \right\}.$$

Several improvements were made in nonparametric theory thanks to this idea. For instance, it is shown in [5] that, for the density estimation model the thresholding procedure is more efficient than the linear procedure, whose

maxiset is given in [15]. And in the heteroscedastic white noise model, [22, 23] shown that thresholding procedures are better than linear estimators and as good as Bayesian procedures. In the case of white noise model, we recall the following result which is a particular case of [22], see [5] for the equivalent result in terms of density estimation.

Proposition 4 *Let $1 \leq p < \infty$, $1 \leq r < \infty$, $s > \frac{d}{r}$ and $\alpha \in (0,1)$. Let $\hat{f}_n^L$ be the linear estimator given in Definition 2. For any $m > 0$ suppose we are given $(\lambda_{j,k}^{(n)}(m))_{j,k}$ weights in $[0,1]$ such that*

- *There exists $c < 1$ such that for all $m_n > 0$ and $j \geq m$, $\lambda_{j,k}^{(n)}(m) \leq c$.*
- *There exists $c_s \in \mathbb{R}$ such that for any $m > 1$*

$$\sum_{1 \leq j \leq m_n} \sum_{k \in \{0,\dots,2^j-1\}^d} (\lambda_{j-1,k}^{(n)}(m) - \lambda_{j,k}^{(n)}(m))(1-\lambda_{j,k}^{(n)}(m))^{p-1} \left(\frac{j}{m}\right)^{-ps} \leq c_s.$$

Then for any $n > 0$, we suppose that we are given $m_n > 0$ such that

- *There exists n such that $m_n \leq 1$,*
- *$n \mapsto m_n$ is continuous,*
- *$m_n \to \infty$ as $n \to \infty$*

We suppose that there exists a positive constant T_1 such that for any $n \in \mathbb{N}$,

$$\frac{m_n^{ps}}{n^{p/2}} \sum_{j,k} \int |\psi_{j,k}|^p \leq T_1. \tag{15}$$

Then for every f, there exists a positive constant C such that for any $n \in \mathbb{N}$,

$$\mathbb{E}\|\hat{f}_n^L - f\|_p^p \leq c m_n^{-sp}$$

if and only if $f \in B_p^{s,\infty}([0,1]^d)$.

Before stating the result associated to thresholding algorithms, we define new function spaces closely related to approximation theory. Those spaces, weak Besov spaces, defined in [5] are subsets of Lorentz spaces, and constitute a larger class than Besov spaces.

Definition 7 *Let $0 < r < p < \infty$. We say that a function $f = \sum_{j,k} c_{j,k}\psi_{j,k}$ belongs to $W(r,p)$ if and only if*

$$\sup_{\lambda > 0} \lambda^r \sum 2^{j(\frac{dp}{2}-d)} \sum_k \mathbb{1}_{\{|c_{j,k}|>\lambda\}} < \infty. \tag{16}$$

A fast calculation shows that the space $W(r,p)$ contains the homogeneous Besov spaces $B_r^{\beta,\infty}$ as soon as $\beta \geq \frac{d}{2}(\frac{p}{r}-1)$.

The maxiset associated with the thresholding estimation procedure is given by a weak Besov space as proved in [5], and developed further in the heteroscedastic regression case in [16].

Proposition 5 *Let $1 \leq p < \infty$, $1 \leq r < \infty$, $s > \frac{d}{r}$ and $\tilde{\alpha} \in (0,1)$.Let $\hat{f}_n^T$ be the estimator defined by (4) and (10). Then for every f we have the following equivalence:*
$\exists K > 0$ such that $\forall n > 0$,

$$\mathbb{E}\|\hat{f}_n^T - f\|_p^p \leq K\left(\sqrt{n\log(n)^{-1}}\right)^{-\tilde{\alpha}p} \tag{17}$$

if and only if $f \in B_p^{\tilde{\alpha}/2,\infty} \cap W((1-\tilde{\alpha})p,p)$.

Furthermore, another important key result involving Besov spaces is the following proposition from [10].

Proposition 6 *Let us define the scaling function of a distribution f by*

$$\forall p > 0 \quad s_f(p) = \sup\{s:\ f \in B_p^{s,\infty}\}. \tag{18}$$

Let s_0 and p_0 be fixed such that $s_0 - \frac{d}{p_0} > 0$. Outside a Haar-null set in $B_{p_0}^{s_0,\infty}(\mathbb{R}^d)$, we have:

$$s_f(p) = \begin{cases} s_0 & \text{if } p \leq p_0 \\ \frac{d}{p} + s_0 - \frac{d}{p_0} & \text{if } p \geq p_0. \end{cases} \tag{19}$$

One can check that a lower bound of this scaling function is given by Besov embeddings and interpolation theory, which can be found in [24]. This result states that one cannot have a better regularity than the one given by those embeddings. In our case, we will exploit this result by comparing those critical spaces with the maxiset associated to each procedure.

4.2. Generic risk for linear estimators

Let $1 \leq p < \infty$, $1 \leq r < \infty$ and $s > \frac{d}{r}$ be fixed. Denote

$$s' = s - \left(\frac{d}{r} - \frac{d}{p}\right)_+,$$

and

$$\alpha(s') = \frac{s'}{2s'+d}.$$

In this section, we prove the first part of Theorem 1. We define the linear estimator as in Definitions 2 and 3. For the sake of simplicity, we assume here that we take projections weights in these two definitions. Let $\theta > 0$ be given. As we are looking for the polynomial behavior of linear estimators, we take $T_n = m_n$ and $m_n = n^\theta$. Define $j_1(n)$ be such that $2^{j_1(n)} \leq m_n < 2^{j_1(n)+1}$. From [7] we know that there exists $c > 0$ such that for any $n \in \mathbb{N}$, and for any $f \in B_r^{s,\infty}([0,1]^d)$,

$$\mathbb{E}\|\hat{f}_n^L - f\|_p^p \leq c\left(2^{-j_1(n)sp} + \left(\frac{2^{j_1(n)}}{n}\right)^{p/2}\right). \tag{20}$$

Furthermore, the infimum in (20) is obtained when the two terms are balanced, that is for $\theta_0 = \frac{1}{2s'+d}$. And we obtain

$$\inf_{\hat{f}_n^L \ linear} \mathbb{E}\|\hat{f}_n^L - f\|_p^p \leq cn^{-\alpha(s')p}. \tag{21}$$

Let us now check the lower bound. Choose $m_n = n^\theta$, with $\theta > 0$ given and let $\hat{f}_n^{L,\theta}$ be the corresponding estimator. In a first time, we have to show that for every $\theta > 0$ and $\varepsilon > 0$ fixed, the set

$$\tilde{M}(\theta,\varepsilon) := \left\{f \in B_r^{s,\infty}([0,1]^d);\ \exists c > 0\ \forall n \in \mathbb{N},\ \mathbb{E}(\|\hat{f}_n^{L,\theta} - f\|_{L^p}^p) < cn^{-(\alpha(s')+\varepsilon)p}\right\}$$

is a Borel Haar null set.

Taking into account that $\int |\psi_{j,k}|^p \sim 2^{jd(\frac{p}{2}-1)}$ we see that equation (15) is

satisfied for $\theta \leq \frac{1}{2s'+d}$. We have thus two cases, if $\theta > \frac{1}{2s'+d}$ then

$$\begin{aligned}
\mathbb{E}(\|\hat{f}_n^L - f\|_p^p) &= \mathbb{E}\left(\int \left(\sum_{j\leq j_1(n)}\sum_k \frac{\varepsilon_{j,k}}{\sqrt{n}}\psi_{j,k}(x) + \sum_{j>j_1(n)}\sum_k c_{j,k}\psi_{j,k}\right)^p\right) \\
&\geq c\mathbb{E}\left(\sum_{j\leq j_1(n)}\sum_k (\frac{\varepsilon_{j,k}}{\sqrt{n}})^p \int \psi_{j,k}(x)^p + \sum_{j>j_1(n)}\sum_k |^p c_{j,k}|^p \int \psi_{j,k}^p\right) \\
&\geq c\sum_{j\leq j_1(n)}\sum_k \mathbb{E}((\frac{\varepsilon_{j,k}}{\sqrt{n}})^p) 2^{dj(\frac{p}{2}-1)} \\
&\geq c\sum_{j\leq j_1(n)}\sum_k (\mathbb{E}((\frac{\varepsilon_{j,k}}{\sqrt{n}})^2))^{p/2} 2^{dj(\frac{p}{2}-1)} \\
&\geq cn^{dp\theta/2-p/2} > n^{-\alpha(s')}.
\end{aligned}$$

Where the first inequality comes from the superconcentration property of wavelets whereas the last one is the Hölder inequality [16]. Thus for $\theta > \frac{1}{2s'+d}$ and for any $\varepsilon > 0$, the set $\tilde{M}(\theta, \varepsilon)$ is an empty set.

Let us treat the case $\theta < \frac{1}{2s'+d}$. From Proposition 4, $\tilde{M}(\theta, \varepsilon)$ is then included in $B_p^{s'+\varepsilon,\infty}([0,1]^d)$. And from Proposition 6, we know that this set is a Haar null Borel set of $B_r^{s,\infty}([0,1]^d)$.

We thus obtain that $\forall\theta > 0$ and $\forall\varepsilon > 0$, the set

$$\left\{f \in B_r^{s,\infty}([0,1]^d);\ \exists c > 0\ \forall n \in \mathbb{N},\ \mathbb{E}(\|\hat{f}_n^{L,\theta} - f\|_{L^p}^p) < cn^{-(\alpha(s')+\varepsilon)p}\right\}$$

is a Haar null set.

This set can also be written,

$$\left\{f \in B_r^{s,\infty}([0,1]^d);\ \limsup_{n\to\infty} \frac{\log(\mathbb{E}(\|\hat{f}_n^{L,\theta} - f\|_{L^p}^p))}{-p\log n} > \alpha(s') + \varepsilon\right\}.$$

Taking the countable union of those sets over a dense sequence θ_n and a decreasing sequence $\varepsilon_n \to 0$, and the complementary we obtain that for almost every function in $B_r^{s,\infty}([0,1]^d)$,

$$\liminf_{n\to\infty} \frac{\log(\mathbb{E}(\|\hat{f}_n^L - f\|_{L^p}^p))}{-p\log n} \leq \alpha(s').$$

Which induces the expected result.

4.3. Thresholding algorithms

In this part, we take the estimation procedures given in Definition 4 and in Definition 5.

Let us turn out our attention to the minimax rate of convergence for this estimator. For this purpose, we write in the following

$$\tilde{\alpha}(s) = \begin{cases} \frac{2s}{2s+d} & \text{if } r > \frac{pd}{2s+d} \\ \frac{2(s-\frac{d}{r}+\frac{d}{p})}{2(s-\frac{d}{r})+d} & \text{else.} \end{cases} \tag{22}$$

The proof of the second point of Theorem 1 follows the same scheme as the previous one. In this case, the upper bound is given in [9]. Thus we know that for every function in $B_r^{s,\infty}([0,1]^d)$, and for all $1 < p < \infty$,

$$\mathbb{E}(\|\hat{f}_n^T - f\|_{L^p}^p) < c\sqrt{\frac{n}{\log n}}^{-\tilde{\alpha}(s)p}.$$

In order to prove the lower bound, we use Proposition 5.

For every values of $\tilde{\alpha}$, let $0 < \varepsilon < 1 - \tilde{\alpha}$ be fixed, and $M(\varepsilon)$ be the set defined by

$$M(\varepsilon) = \left\{ f \in B_r^{s,\infty}([0,1]^d);\ \exists c > 0\ \forall n \in \mathbb{N},\ \mathbb{E}(\|\hat{f}_n^T - f\|_{L^p}^p) < c\sqrt{\frac{n}{\log n}}^{-(\tilde{\alpha}(s)+\varepsilon)p} \right\}.$$

Thanks to Proposition 5, this set $M(\varepsilon)$ is embedded in $B_p^{\frac{\tilde{\alpha}+\varepsilon}{2},\infty} \cap W((1-\tilde{\alpha}-\varepsilon)p, p)$.

The end of the proof is based on the following proposition.

Proposition 7 *Let f be a given distribution. Let us define the weak scaling function of a distribution f by*

$$\forall p > 0 \quad \tilde{s}_f(p) = \sup\{\alpha :\ f \in W((1-\alpha)p, p)\}. \tag{23}$$

Let s and r be fixed such that $s - \frac{d}{r} > 0$. Outside a Haar-null set in $B_r^{s,\infty}(\mathbb{R}^d)$, we have:

$$\tilde{s}_f(p) = \begin{cases} \frac{2s}{2s+d} & \textit{if } r > \frac{pd}{2s+d} \\ \frac{2(s-\frac{d}{r}+\frac{d}{p})}{2(s-\frac{d}{r})+d} & \textit{else.} \end{cases} \tag{24}$$

Proof: In order to prove Proposition 7, let us prove that $W((1-\tilde{\alpha}-\varepsilon)p,p)$ is a Haar null Borel set in $B_r^{s,\infty}([0,1]^d)$. For this purpose, we define our transverse measure as the probe generated by the function g defined by its wavelet coefficients:

$$d_{j,k} = \frac{2^{-(s-\frac{d}{r}+\frac{d}{2})j} 2^{-\frac{d}{r}J}}{j^a}$$

where $a = 1+\frac{3}{r}$ and $0 \leq J \leq j$ and $K \in \{0,\dots,2^J-1\}^d$ are such that

$$\frac{K}{2^J} = \frac{k}{2^j}$$

is an irreducible fraction. As it can be seen in Proposition 2 of [14], this function g belongs to $B_r^{s,\infty}([0,1]^d)$. Let $f \in B_r^{s,\infty}([0,1]^d)$ be an arbitrary function and consider the affine subset

$$M = \{\alpha \in \mathbb{R} \;\; f+\alpha g \in W((1-\tilde{\alpha}-\varepsilon)p,p)\}.$$

Suppose that there exist two points α_1 and α_2 in M. Thus $f+\alpha_1 g-(f+\alpha_2 g)$ belongs to $W((1-\tilde{\alpha}-\varepsilon)p,p)$, and there exists $c>0$ such that

$$\|f+\alpha_1 g-(f+\alpha_2 g)\|_{W((1-\tilde{\alpha}-\varepsilon)p,p)} = \|(\alpha_1-\alpha_2)g\|_{W((1-\tilde{\alpha}-\varepsilon)p,p)} \leq c. \quad (25)$$

As a fast calculation shows that

$$\forall \alpha > 0, \;\; \|\alpha g\|_{W(r,p)} = \alpha^r \|g\|_{W(r,p)} \quad (26)$$

we just have now to determine $\|g\|_{W(r,p)}$. Thanks to equation (16), this is equivalent to determine for every $t>0$ the value of

$$2^{-(1-\tilde{\alpha}-\varepsilon)pt} \sum_{j\geq 0} 2^{j(\frac{dp}{2}-d)} \sum_k \mathbb{1}_{\{d_{j,k}>2^{-t}\}}$$

But by definition of g, we have,

$$\frac{2^{-(s-\frac{d}{r}+\frac{d}{2})j} 2^{-\frac{d}{r}J}}{j^a} > 2^{-t} \Rightarrow (s-\frac{d}{r}+\frac{d}{2})j + \frac{d}{r}J \leq t,$$

which implies that

$$J \leq \frac{r}{d}t - (s-\frac{d}{r}+\frac{d}{2})\frac{r}{d}j.$$

Note that the condition $J \geq 0$ implies also that j is limited by

$$j(s - \frac{d}{r} + \frac{d}{2}) \leq t.$$

We denote by $\tilde{t} = \frac{t}{s-\frac{d}{r}+\frac{d}{2}}$ and by $\tilde{\tilde{t}} = \frac{t}{s+\frac{d}{2}}$. Thus we have, for every $t > 0$,

$$\|g\|_{W((1-\tilde{\alpha}-\varepsilon)p,p)} \geq 2^{-(1-\tilde{\alpha}-\varepsilon)pt} \sup_{0\leq j\leq \tilde{t}} 2^{j(\frac{dp}{2}-d)} \sum_{J=0}^{j\wedge[\frac{r}{d}t-(s-\frac{d}{r}+\frac{d}{2})\frac{r}{d}j]} 2^{dJ}$$

$$\geq 2^{-(1-\tilde{\alpha}-\varepsilon)pt} \sup\left(\sup_{0\leq j\leq \frac{t}{s+\frac{d}{2}}} 2^{j(\frac{dp}{2}-d)} \sum_{J=0}^{j} 2^{dJ}, \sup_{\frac{t}{s+\frac{d}{2}}+1\leq j\leq \tilde{t}} 2^{j(\frac{dp}{2}-d)} \sum_{J=0}^{[\frac{r}{d}t-(s-\frac{d}{r}+\frac{d}{2})\frac{r}{d}j]} 2^{dJ} \right)$$

$$\geq \frac{2^{-(1-\tilde{\alpha}-\varepsilon)pt}}{2^d-1} \sup\left(\sup_{0\leq j\leq \tilde{\tilde{t}}} 2^{\frac{dpj}{2}}(1-2^{-jd}), \sup_{\tilde{\tilde{t}}<j\leq \tilde{t}} 2^{j(\frac{dp}{2}-d)}(2^{rt}2^{-jr(s+\frac{d}{2}-\frac{d}{r})}-1) \right)$$

Merging this result with (25) together with (26), we obtain that, if there exist α_1 and α_2 in M then they satisfy that for every $t \geq 0$ and $0 \leq j \leq \tilde{t}$,

$$|\alpha_1 - \alpha_2|^{(1-\tilde{\alpha}-\varepsilon)p} \leq \inf\left(\frac{c2^{(1-\tilde{\alpha}-\varepsilon)pt}}{\sup_{0\leq j\leq \tilde{\tilde{t}}} 2^{\frac{dpj}{2}}|1-2^{-jd}|}, \frac{c2^{(1-\tilde{\alpha}-\varepsilon)pt}}{\sup_{\tilde{\tilde{t}}<j\leq \tilde{t}} 2^{j(\frac{dp}{2}-d)}|2^{rt}2^{-jr(s+\frac{d}{2}-\frac{d}{r})}-1|} \right) \tag{27}$$

We have thus two cases:

- If $r > \frac{dp}{2s+d}$

  $$\tilde{\alpha} = \frac{2s}{2s+d}.$$

  But, if we take the first term,

  $$\sup_{0\leq j\leq \tilde{\tilde{t}}} 2^{\frac{dpj}{2}}|1-2^{-jd}| \sim 2^{\frac{tdp}{2s+d}},$$

  we have

  $$|\alpha_1 - \alpha_2|^{(1-\tilde{\alpha}-\varepsilon)p} \leq c2^{-\varepsilon pt}. \tag{28}$$

- When $r \leq \frac{dp}{2s+d}$, and as $s > \frac{d}{r}$ we have necessarily $p > 2$ and we obtain

  $$\tilde{\alpha} = \frac{2(s-\frac{d}{r}+\frac{d}{p})}{2(s-\frac{d}{r})+d}.$$

In this case,

$$\sup_{\tilde{\tilde{t}}<j\leq\tilde{t}} 2^{j(\frac{dp}{2}-d)}|2^{rt}2^{-jr(s+\frac{d}{2}-\frac{d}{r})}-1| \sim 2^{\frac{td(p-2)}{2(s-\frac{d}{r})+d}}.$$

And once again,

$$\forall t>0 \quad |\alpha_1-\alpha_2|^{(1-\tilde{\alpha}-\varepsilon)p} \leq c2^{-\varepsilon pt}. \tag{29}$$

As $1-\tilde{\alpha}-\varepsilon>0$, it can be deduced from equations (28) and (29) that for t large enough, M is of vanishing Lebesgue measure and $W((1-\tilde{\alpha}-\varepsilon)p,p)$ is an Haar null set in $B_r^{s,\infty}([0,1]^d)$. □

Thanks to invariance under inclusion, we have obtained that for every $\varepsilon>0$, the set of functions f in $B_r^{s,\infty}([0,1]^d)$ such that

$$\exists c>0\ \forall n\in\mathbb{N},\ \mathbb{E}(\|\hat{f}_n^T-f\|_{L^p}^p)<c\sqrt{\frac{n}{\log n}}^{-(\alpha(s)+\varepsilon)p}$$

is a Haar null set.

The end of the proof is similar to part 4.2.

References

[1] F. Autin, *Point de vue maxiset en estimation non paramétrique*, Ph.D. thesis, Université Paris 7, 2004.

[2] Y. Benyamini and J. Lindenstrauss, *Geometric nonlinear functional analysis. Volume 1*, Colloquium Publications. American Mathematical Society (AMS), 2000.

[3] L. Birgé, *Approximation dans les espaces métriques et théorie de l'estimation*, Z. Wahrsch. Verw. Gebiete **65** (1983), no. 2, 181–237.

[4] J.P.R. Christensen, *On sets of Haar measure zero in Abelian Polish groups*, Israel J. Math. **13** (1972), 255–260.

[5] A. Cohen, R. DeVore, G. Kerkyacharian, and D. Picard, *Maximal spaces with given rate of convergence for thresholding algorithms*, Appl. Comput. Harmon. Anal. **11** (2001), no. 2, 167–191.

[6] I. Daubechies, *Orthonormal bases of compactly supported wavelets*, Comm. Pure and App. Math. **41** (1988), 909–996.

[7] D. Donoho and I. Johnstone, *Minimax risk over l_p-balls for l_q-error*, Probab. Theory Related Fields **99** (1994), no. 2, 277–303.

[8] ———, *Minimax estimation via wavelet shrinkage*, Ann. Statist. **26** (1998), no. 3, 879–921.
[9] D. L. Donoho, I. M. Johnstone, G. Kerkyacharian, and D. Picard, *Universal near minimaxity of wavelet shrinkage*, Festschrift for Lucien Le Cam, Springer, New York, 1997, pp. 183–218.
[10] A. Fraysse, *Generic validity of the multifractal formalism*, SIAM J. Math. Anal. **37** (2007), no. 2, 593–607.
[11] B. Hunt, *The prevalence of continuous nowhere differentiable function*, Proceed. A.M.S **122** (1994), no. 3, 711–717.
[12] B. Hunt, T. Sauer, and J. Yorke, *Prevalence: A translation invariant "almost every" on infinite dimensional spaces*, Bull. A.M.S **27** (1992), no. 2, 217–238.
[13] I. A. Ibragimov and R. Z. Has'minskiĭ, *Statistical estimation*, Applications of Mathematics, vol. 16, Springer-Verlag, 1981.
[14] S. Jaffard, *On the Frisch-Parisi conjecture*, J. Math. Pures Appl **79** (2000), 525–552.
[15] G. Kerkyacharian and D. Picard, *Density estimation by kernel and wavelets methods: optimality of Besov spaces*, Statist. Probab. Lett. **18** (1993), no. 4, 327–336.
[16] ———, *Thresholding algorithms, maxisets and well-concentrated bases*, Test **9** (2000), no. 2, 283–344, With comments, and a rejoinder by the authors.
[17] ———, *Minimax or maxisets?*, Bernoulli **8** (2002), no. 2, 219–253.
[18] S. Mallat, *A wavelet tour of signal processing*, San Diego, CA: Academic Press. xxiv, 1998.
[19] Y. Meyer, *Ondelettes et opérateurs*, Hermann, 1990.
[20] A. S. Nemirovskiĭ, B. T. Polyak, and A. B. Tsybakov, *The rate of convergence of nonparametric estimates of maximum likelihood type*, Problemy Peredachi Informatsii **21** (1985), no. 4, 17–33.
[21] M. S. Pinsker, *Optimal filtration of square-integrable signals in Gaussian noise*, Problems Inform. Transmission **16** (1980), no. 2, 52–68.
[22] V. Rivoirard, *Maxisets for linear procedures*, Statist. Probab. Lett. **67** (2004), no. 3, 267–275.
[23] ———, *Nonlinear estimation over weak Besov spaces and minimax Bayes method*, Bernoulli **12** (2006), no. 4, 609–632.
[24] E. Stein, *Singular integrals and differentiability properties of functions*, Princeton University Press, 1970.
[25] A. Van der Vaart, *Asymptotic statistics*, Cambridge Series in Statisti-

cal and Probabilistic Mathematics, vol. 3, Cambridge University Press, 1998.